Inference-Based Biodatabase Integration as a Tool for Microarray Analytics
Goertzel, Ben
Biomind LLC, Wheaton, MD, USA

The BiomindDB is a unique integrative database, created at Biomind LLC, formed by applying probabilistic inference algorithms to an integrated knowledge base formed by amalgamating numerous public biology databases, including the Gene Ontology, PIR, GenBank, SGD, Kegg, Unigene and MGI, as well as a large number of gene expression datasets and pertinent subsets of chemical databases. Initially the BiomindDB will be browsable via a simple Web UI; a virtual-reality-based graphical navigation interface is also under development.

In addition to its intrinsic value, the BiomindDB is intended specifically for use within the data analysis process. We have developed tools for analyzing microarray gene and protein expression data, which have the ability to integrate biological background knowledge into the data analysis process itself. In these tools, knowledge from the BiomindDB is used inside algorithms carrying out analytical tasks such as clustering, supervised categorization, and time series analysis. For example, incorporation of background knowledge into the analytic process may reveal complex regulatory relationships that would otherwise go undetected.

Among other applications, this technology is currently being applied to the analysis of gene expression data gathered by the Center for Disease Control, regarding individuals with Chronic Fatigue Syndrome. The application here is to find diagnostic rules predicting whether an individual has CFS or not, based on the expression levels of genes in peripheral blood mononuclear cells. The use of BiomindDB knowledge within the analytic process allows the automated learning of rules that involve the expression of functional categories of genes as well as gene expression levels directly.